



المقالة الاستراتيجية

التاريخ التحريري: 24 أكتوبر 2026

الحاسوب الذي يخفض أجره بنفسه

لماذا قد يسبق سقوط سعر الذكاء استرداد ثمنه؟

The Token Deflation Trap: When Faster Inference Undermines Compute Payback

استهلال: قد يكون المعالج ممتازًا في أداء المهمة التي اشترى لها، ومتأخرًا اقتصاديًا في اليوم نفسه. هذه ليست أحجية عن أعطال الرقائق، بل عن صناعة تستطيع أن تزيد ما تنتجه من الذكاء بسرعة تفوق قدرتها على تقاضي ثمنه. نفتح ملفًا ماليًا أمريكيًا، وننتقل إلى خوارزمية تحاكي صفحات الذاكرة، ثم إلى حوسبة دخلت التشغيل في السعودية؛ لنكتشف أن المنافسة الحاسمة ربما تدور حول الدقيقة التي يصبح فيها الحساب مفيدًا وقابلًا للبيع، لا حول عدد الحسابات الممكنة على الورق.

1. في دفاتر المحاسبين تعيش الرقاقة ست سنوات؛ وفي السوق تولد لها منافسة كل فصل

في التقرير السنوي لشركة «كور ويف» (CoreWeave) لعام 2025، توجد جملة صغيرة يكاد القارئ يعبرها مثل موظف لم يجد في كشف الراتب مفاجأة: منذ يناير 2023 عدلت الشركة العمر الإنتاجي المقدّر لمعدات الحوسبة في مراكز بياناتها من خمس سنوات إلى ست. سبب التعديل، كما ذكر إفصاحها، يشمل تقدم الأجهزة وتحسين البرمجيات وتصميم مراكز البيانات. هذا عمر محاسبي مقدّر يُستخدم في توزيع الإهلاك (Depreciation)، وليس وعدًا بأن الجهاز سيبقى أعلى الأجهزة ربحية ست سنوات S01.

في 5 يناير 2026، قدّمت شركة «إنفيديا» (NVIDIA) منصتها «روبين» (Rubin)، معلنة أن تصميمها المتكامل قد يخفض تكلفة إنتاج رمز الاستدلال (Inference Token Cost) بما يصل إلى عشرة أمثال مقارنة بمنصة «بلاكويل» (Blackwell) في أعباء عمل محددة. إعلان من الجهة البائعة، يصف قدرة تقنية مستهدفة؛ ليس كشفًا بأسعار بيع فعلية، ولا حكمًا على جميع النماذج والتطبيقات S03.

ثم جاء إفصاح النصف الأول من 2026: قالت «كور ويف» (CoreWeave) إن التزامات الأداء المتبقية (Remaining Performance Obligations; RPO) في نهاية يونيو بلغت 103.7 مليارات دولار. الرقم مدهش، والأهم تعريفه؛ إنه قيمة تعاقدية موزعة على خدمات لم تُقدّم بعد، وفق قواعد حسابية وتأثيرات محتملة للتأخير والخصومات، وليس نقدًا في الخزنة ولا صافي أرباح مؤكدًا. توقعت الشركة الاعتراف بنحو 41% منه خلال الأربعة والعشرين شهرًا التالية، ثم 39% في الأربعة والعشرين التي تليها S02.

من المعتاد أن يبدأ الحديث عن الحوسبة الاصطناعية باسم الشريحة وسعة ذاكرتها. لكن ملأً يحوي «ست سنوات» و«خفض تكلفة الرمز عشرة أضعاف» و«103.7 مليارات من الخدمات المستقبلية» يضعنا أمام مسألة أظرف وأخطر: **أي زمن من هذه الأزمنة يجب أن يربح السباق؟ زمن الإهلاك؟ زمن الجيل التقني الجديد؟ زمن العقود التي تحمي التدفقات النقدية؟ أم زمن تغير السعر الذي يقبل العميل دفعه؟**

لم يثبت أيٌّ من الأرقام انهيار سوق، ولا يثبت أن شركة بعينها عالققة في فخ. على العكس، وجود العقود طويلة الأجل سبب مهم لثلا نقع في حكاية «معالج يصبح بلا قيمة فور صدور خلفه». لكنه يثبت أن السؤال الصحيح عن الاستثمار ليس «هل تعمل الرقاقة؟»، بل «من سيشتري ساعات عملها، وبأي سعر، وكما مرة، وقبل أي موعد لاستحقاق رأس المال؟».

2. الخبر الذي يصل إلى المطور بوصفه تخفيضًا، قد يصل إلى الممول بوصفه فرضية جديدة

في سبتمبر 2026 أعلنت «أوبن إيه آي» (OpenAI) أن أسعار واجهتي النماذج (Application Programming Interfaces; APIs) لـ «جي بي تي 6 سول» (GPT-6 Sol) و«جي بي تي 6 لونا» (GPT-6 Luna) انخفضت عما يقابلها في التسعير الترويجي للجيل السابق، وعزت ذلك جزئيًا إلى تحسين التخزين المؤقت والاستدلال. في المثال المنشور لمدخلات «سول»، هبط السعر من 4 دولارات إلى دولارين لكل مليون رمز، وللمخرجات من 20 إلى 10 دولارات. هذه مقارنة بين منتجين من جيلين مختلفين وشروط تسعير معلنة، لا إثباتًا بأن الإيجار السوقي لكل شريحة انخفض إلى النصف S04.

من يقف خلف تطبيق صغير قد يرى في هذا النبأ مساحة للتجريب كانت تضيقها فاتورة الاستدلال: اختبار عشر صيغ بدل خمس، أو خدمة مستخدم لم يكن سعر المحادثة معه يسمح بنموذج أكثر قدرة. أما من اشترى رفوف المعالجات وحوّل ثمنها إلى التزام تمويلي طويل، فسيطرح سؤالاً آخر: هل سيؤدي انخفاض سعر الخدمة إلى طلب يتضاعف بالسرعة نفسها، أم سيحتاج إلى بيع عمل أكثر مقابل الإيراد ذاته؟

هنا يولد ما أسميه **فخ انكماش سعر الرمز (Token Deflation Trap)**. لا يقع الفخ لمجرد أن السعر هبط. يقع إذا تزامن ثلاثة أمور: انخفاض الإيراد لكل وحدة عمل تُباع، وتعذر رفع عدد الوحدات المفيدة والمدفوعة بالقدر اللازم، واستحقاق رأس مال لا تنخفض كلفته بالسرعة ذاتها. قد تفشل واحدة من هذه الشروط فينقلب الفخ فرصة ذهبية. لذلك لا يصلح الاسم نبوءة، بل يصلح سؤالاً يُمسك بالمفاصل.

والسوق أشد تعقيدًا من «سعر لكل مليون رمز». فالرمز الداخل إلى النموذج ليس الرمز الخارج منه، والرمز الذي أُعيد استخدام سياقه المحفوظ (Cached Input Token) ليس حسابًا جديدًا، والجواب المختصر الناجح قد يغني عن سلسلة طويلة من المحاولات. إذا باعنا «مليون رمز» كأنا نبيع مليون مسمار متماثل، سنكتشف أن بعض المسامير يستهلك ذاكرة، وبعضها يحتاج زمن انتظار، وبعضها خرج من صندوق التخزين المؤقت أصلًا. الوحدة الحسابية مريحة؛ الوحدة الفيزيائية غير مطيعة.

3. رحلة الرمز التي لا تراها شاشة المحادثة

لنفهم لماذا يستطيع البرنامج أن يعيد تسعير الحديد، لندخل من الجملة التي يكتبها مستخدم عادي: «لخص هذه المراسلات واربط القرار المالي بمراسلات الأسبوع الماضي». لا تتلقى وحدة المعالجة الرسومية (Graphics Processing Unit; GPU) جملة جاهزة للتفكير، بل سلسلة من الرموز (Tokens) تمر عبر عمليات حسابية مختلفة في سلوكها نحو الذاكرة والنقل والتوازي.

في مرحلة ملء السياق الأولية (Prefill)، تُعالج الرموز الواردة بما يسمح ببناء تمثيلات الانتباه (Attention). ثم تبدأ مرحلة التوليد المتكرر (Decode): يُنتج رمز، ثم يُستخدم ضمن سياق الرمز التالي. الأولى تميل، بحسب حجم الحمل والمعمارية، إلى كثافة حسابية قابلة للتوازي؛ والثانية قد تصطدم بعرض نطاق الذاكرة (Memory Bandwidth) وحركة الأوزان وحالة المحادثة. ليس من الضروري أن تكون الشريحة الممتازة في الأولى هي صاحبة أفضل تكلفة نهائية في الثانية. لهذا توثق أدوات مثل «تينسور آر تي إل إل إم» (TensorRT-LLM) هندسات تفصل موارد المرحلتين عند الحاجة بدل إجبارهما على إيقاع تشغيل واحد S07.

وللمحادثة ذاكرة عمل لا تشبه الحفظ الطويل للمعلومات: مخبأ المفتاح والقيمة (Key-Value Cache; KV Cache) يحتفظ بتمثيلات مفيدة لمنع إعادة حساب أجزاء من السياق. كل محادثة طويلة تطلب مكانًا إضافيًا. إذا حُجز لها مكان واسع ثم استخدمت جزءًا صغيرًا، فقد يتكدس داخل شريحة غالية فراغ لا يستطيع أحد دفع فاتورته. إذا نظرت إلى الذاكرة كما ينظر أمين مكتبة إلى الرفوف، ستعرف أن المشكلة ليست دائمًا قلة الكتب؛ قد تكون الطريقة التي نحجز بها الأدراج.

عام 2023 نشر باحثون نظام «في إل إل إم» (vLLM) وخوارزمية «الانتباه المرقّم بالصفحات» (PagedAttention)، المستوحاة من ذاكرة الصفحات الافتراضية (Virtual Memory Paging) في أنظمة التشغيل. عوض الحجز المتصل المهدر، قسّموا مخبأ الانتباه إلى كتل تُدار وتُشارك بمرونة. في تجارب الورقة المحكمة، تحسّن معدل إنتاج الاستدلال بنحو ضعفين إلى أربعة أضعاف مقارنة بأنظمة مرجعية محددة مع مستويات تأخير متقاربة S05. لا يعني هذا أن كل مركز بيانات سيحصل على التحسن ذاته بمجرد تثبيت برنامج، لكنه يثبت إمكانية أن يعيد قرار خوارزمي توزيع قيمة العتاد الموجود.

وقبلها بيّن نظام «أوركا» (Orca) في مؤتمر نظم التشغيل (OSDI 2022) أن جدولة العمل عند مستوى دورة توليد واحدة (Iteration-Level Scheduling) يمكن أن تخفف انتظار الطلبات خلف دفعات أطول منها S06. تقنيات الدفعات المستمرة (Continuous Batching) وإعادة استخدام مقدمة الطلب (Prefix Caching) وفصل مرحلتي الحساب ليست ملحقات تجميلية؛ إنها طرق لتغيير مقدار العمل الذي يتحول إلى خدمة ضمن زمن مقبول. أما التكميم (Quantization) فيقلل دقة التمثيل العددي وأثره على الذاكرة والحساب، لكنه يحتاج اختبارات جودة لأن الوفر قد يدفع ثمنه المستخدم في دقة الإجابة. والتوليد التخميني (Speculative Decoding) يقترح رموزًا بدالة أسرع ثم يتحقق منها نموذج آخر، فلا تكون نسبة التسريع ثابتة عبر الأحمال.

لهذا لا يجوز مساواة الذروة النظرية لعمليات الفاصلة العائمة (Floating-Point Operations per Second; FLOPS) مع عدد الأجوبة الصحيحة التي يشتريها عميل في الساعة. بين الرقمين طابور، ونموذج، ومُجدول، و طاقة، وعقد سرعة استجابة، وأسئلة قد تطول أو تقصر. الكفاءة ليست ملكية ثابتة للشريحة؛ إنها سلوك منظومة كاملة وهي تنفذ عملاً بعينه.

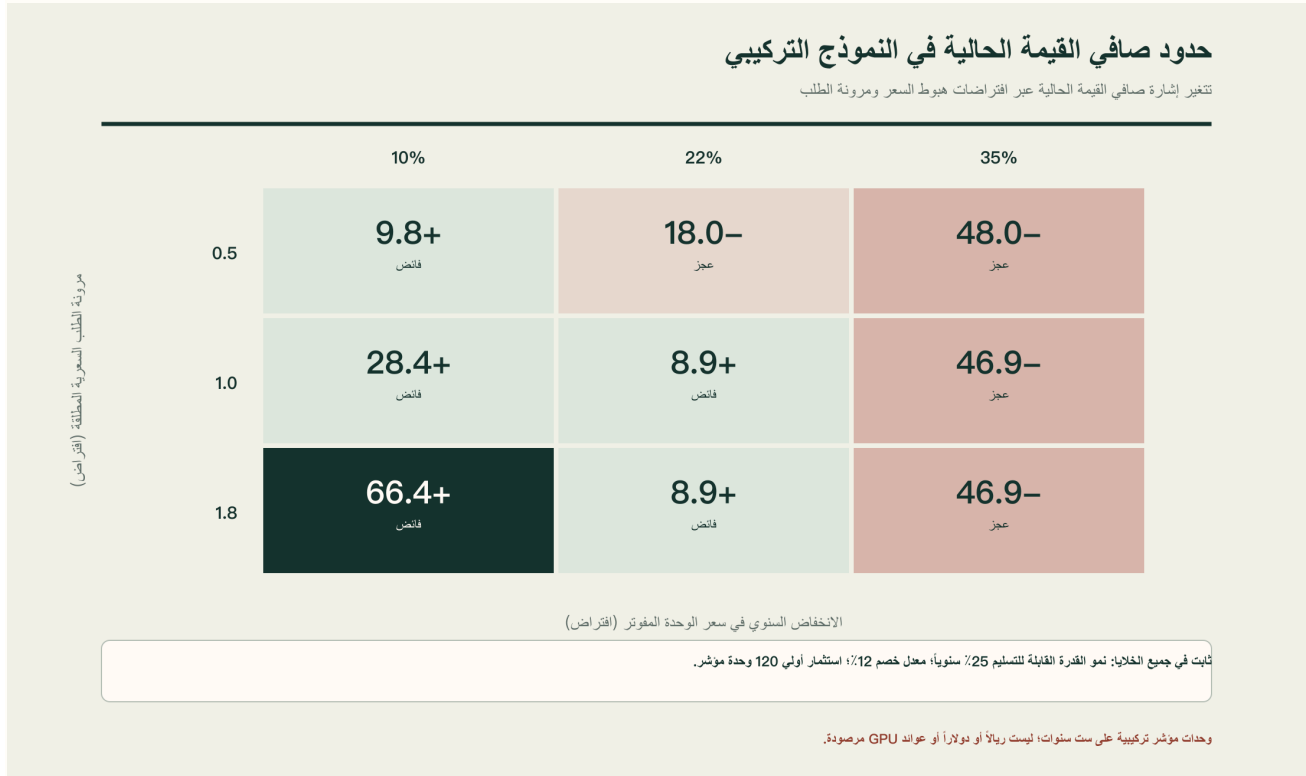
4. المفارقة: يستطيع البرنامج أن ينقذ الاستثمار، وأن يضغط سعره في الوقت نفسه

خذ مثلاً افتراضياً، ولا تنسب أرقامه إلى السعودية أو أي شركة: كان خادم يستجيب لخمسين ألف وحدة عمل في اليوم. بعد تحسين إدارة الذاكرة ارتفعت قدرته إلى ثمانين ألفاً من دون استبداله. هذه زيادة إنتاجية جميلة. لكن إذا استطاع جميع المنافسين نشر التحسين نفسه، تقل ندرة الحساب وتدخل الأسعار في مفاوضات جديدة. يستفيد المستخدم مرتين، مرة من السرعة ومرة من المنافسة؛ ولا يضمن صاحب الخادم أن يحصل هو على المنفعتين.

يتوقف الحكم على ما يسميه الاقتصاديون مرونة الطلب السعرية (Price Elasticity of Demand): كم تزيد الكمية المطلوبة عندما ينخفض السعر؟ في صورة مبسطة، إذا كان الطلب يتناسب مع السعر مرفوعاً لقوة سالبة مقدارها ε ، فإن الإيراد، قبل قيود السعة والتكاليف، يتناسب مع السعر مرفوعاً للقوة $\varepsilon - 1$. عندما تكون ε أعلى من واحد، قد ترتفع الإيرادات الإجمالية مع انخفاض السعر؛ وإذا قلت عن واحد، قد تهبط رغم زيادة الاستهلاك. وعندما تكون السعة ممتلئة، لا يهم الحماس الحسابي للطلب إن لم نستطع تسليم وحداته المدفوعة.

هذه نسخة دقيقة من السؤال الذي طرحه وليام ستانلي جيفونز (William Stanley Jevons) في كتاب «مسألة الفحم» (The Coal Question) عام 1865: أيمن للكفاءة أن تزيد الاستهلاك الإجمالي للموارد بدل أن تنقصه؟ لم يكن يتنبأ بوحدات الذكاء الاصطناعي. والأهم أننا لا نملك رخصة لتحويل ملاحظته التاريخية إلى شعار «كل تحسن يضاعف الطلب حتماً». أثر الارتداد (Rebound Effect) احتمال تحدده المرونة، والقدرة المتاحة، وسعر التطبيقات البديلة، وقيمة الوظيفة المنجزة. إن أراد القارئ قاعدة واحدة من هذا التاريخ فلتكن: لا تستنتج منحنى الطلب من شدة إعجاب المهندس بالمعالج S08.

لذلك يستحق السوق سؤالين متعاكسين، لا إعلانًا واحدًا: هل تؤخر البرمجيات الحاجة إلى شراء الجيل القادم، فترفع عائد المعدات الموجودة؟ وهل تجعل تلك البرمجيات الخدمة وفيرة إلى درجة تُخفض سعر بيعها على الجميع؟ الإجابتان قد تكونان «نعم» في الوقت نفسه، وتحديد الربح لا يتم بعدد الرموز التي خرجت من الخادم بل بالهامش النقدي الذي بقي بعد بيعها.



خريطة العائد وفق المرونة والهبوط السعري، محاكاة توضيحية وليست بيانات شركة

الشكل 1. قيمة حالية صافية (Net Present Value; NPV) مقاسة بوحدات افتراضية، لا بالدولار أو الريال، ثلاث درجات هبوط سعري وثلاث مروّنات طلب، مع تحسن سنوي مفترض في السعة القابلة للبيع بنسبة 25%. كل خلية ناتجة عن تشغيل النموذج المرفق، وليست توقّعاً عن السوق.

5. حين يفسد مؤشر الإشغال فهم المال

ينظر المستثمر أحياناً إلى نسبة انشغال المعالجات (GPU Utilization) كما ينظر صاحب فندق إلى الغرف الممتلئة. التشبيه مفيد إلى أن نتذكر أن الغرفة في الفندق لا تتحول، في أثناء الليل، من جناح فاخر إلى كرسي على ممر بسبب تحديث برمجي. وليس كل تشغيل لمعالج حوسبة قابلة للفوترة: قد يُستهلك في إعداد البيانات، وإعادة محاولة فاشلة، وانتظار بين الشبكة والذاكرة، وعمل داخلي لا يشتريه العميل، أو تنفيذ بلا جودة كافية لتسليمه.

يمكن أن تبدو وحدة معالجة مشغولة بنسبة 95% بينما تحقق إيراداً أقل من وحدة مشغولة بنسبة 65% إذا اختلف السعر، أو قيمة الطلبات، أو مقدار الهدر غير المفوتر، أو مستوى الخدمة. على العكس، قد تنخفض نسبة الإشغال بعد تحسين برمجي ناجح لأن السعة الحقيقية ارتفعت أسرع من الطلب، ولا تكون النسبة المتراجعة علامة فشل تشغيلي.

إنها مسألة مقام الكسر، لا قصة أخلاقية عن الكسل والنشاط.

في النموذج الحسابي المرفق، تبدأ قدرة افتراضية تساوي خمسين وحدة خدمة سنوياً وسعر الوحدة واحداً. نفترض تكلفة رأسمالية تعادل 120 وحدة محاسبية، ونافذة ست سنوات، وخصماً سنوياً قدره 12%، وتكلفة تشغيل ثابتة ومتغيرة معلنة. لا تستند هذه القيم إلى حسابات شركة بعينها؛ قيمتها أنها تتيح تكذيب الحدس بسهولة لا تستطيعها البلاغة.

عندما يهبط السعر المفترض 22% في كل سنة، وتتحسن السعة الفعالة 25%، وتبلغ مرونة الطلب 0.5، تخرج القيمة الحالية الصافية سالبة بنحو 18 وحدة. عند المرونة 1.0 تصبح موجبة بنحو 8.9 وحدات. أما رفع المرونة إلى 1.8 فلا يرفع الناتج في هذا الزوج من الافتراضات: يبقى 8.9 تقريباً، لأن الطلبين يصطدمان بسقف القدرة القابلة للتسليم S12. إذا ازدحم السوق بالطلبات في الخارج، فلا يدفع الطلب الزائد ثمن وحدات لا يستطيع المركز توفيرها.

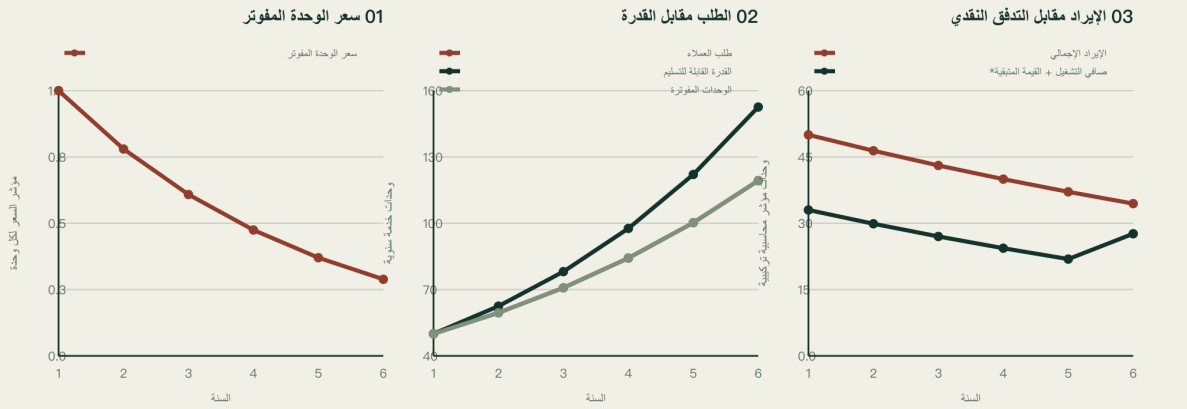
وشغلنا شبكة من 27 سيناريو تجمع ثلاثة معدلات لانخفاض السعر، وثلاث مروونات، وثلاث سرعات لتحسن الطاقة الإنتاجية. خرجت 12 حالة بقيمة صافية موجبة و15 سالبة. ليس في العدد إحصاء عن احتمال الربح؛ جميع الخانات حالات تركيبية اختيرت مسبقاً للاختبار. الدرس الأقوى هو أن حكماً جذاباً مثل «الذكاء الرخيص سيأكل عائد الرقائيق» وحكماً معاكساً مثل «ارتفاع الطلب سينقذ كل الاستثمارات» يفشلان كقانونين عامين. كلاهما يحتاج شرطاً محدداً يمكن قياسه S12.

وثمة خدعة أخرى أجمل من خدعة الإشغال: في الحالة الأساسية للمحاكاة تستعيد التدفقات مجموع التكلفة الرأسمالية اسمياً في السنة الخامسة، لكن قيمتها الحالية الصافية تظل سالبة بنحو 5.6 وحدات. ليس تضارباً في الحساب؛ القيمة التي تعود بعد خمس سنوات لا تساوي القيمة التي خرجت اليوم عندما يكون لرأس المال ثمن. وهكذا يستطيع مشروع أن يحتفل بـ«الاسترداد» ويكون، وفق معيار خصم معلن، أقل جاذبية من بديله. لا تكفي صورة عقد طويل لتجيب عن سؤال تكلفة الانتظار.

ومن المهم ألا يختبئ افتراضنا المتفائل: سمحنا بانخفاض التكلفة المتغيرة لكل وحدة مع تحسن السعة بالنسبة نفسها. حين أبقينا تلك التكلفة ثابتة في اختبار حساسية منفصل، مع بقاء بقية افتراضات الحالة الأساسية، هبطت القيمة الصافية من سالب 5.6 إلى سالب 31.2 وحدة. بهذا المعنى لا يستطيع عرض شرائح جميل عن «زيادة الإنتاجية» أن يحل محل قياس استهلاك الطاقة والعمليات الفعلية لكل خدمة. فسرعة إنجاز مزيد من العمل لا تعني دائماً أن كل عمل صار أرخص S12.

فجوة السعر والقدرة

الحالة الأساسية: $\delta=22\%$ ، $\varepsilon=0.7$ ؛ نمو القدرة القابلة للتسليم 25٪ سنوياً



* يتضمن تدفق السنة 6 قيمة متبقية مقترضة قدرها 8 وحدات؛ لا يظهر الاستثمار الأولي 120 وحدة في السنة صفر.

مخرجات نموذج افتراضي: لا توجد قدرة أو أسعار أو تكاليف نقدية لمشروع مرصودة.

حركة السعر والحجم والإيراد خلال ست سنوات، مثال افتراضي

الشكل 2. ثلاث سلاسل من تجربة مالية تركيبية واحدة. تتباعد الحركة الفيزيائية للسعة عن الحركة التجارية للإيراد؛ ويمثل الطلب ما يرغب السوق في ثرائه، لا ما يتم تسليمه فعلياً. حدود الرسم ومعاملاته كاملة في المرافق البحثية.

6. السعودية: من مشروع الرفوف إلى اقتصاد العملاء الحقيقيين

في 31 أغسطس 2026 أعلنت «إيه إم دي» (AMD) و«سيسكو» (Cisco) و«هيومين» (HUMAIN) أن بنية حوسبة في السعودية مبنية على معالجات «إم آي 355 إكس» (AMD Instinct MI355X)، وشبكات «سيليكون ون» (Cisco Silicon One) وبرمجيات «روكم» (ROCm) دخلت التشغيل وتخدم عملاء. وأعلنت الأطراف نفسها خطة لبدء نشر قدرة تصل إلى 250 ميغاواط في 2027، وطموح توسع يصل إلى غيغاواط واحد بحلول 2030. هذه ثلاث حالات مختلفة يجب ألا تُكتب بصيغة واحدة: تشغيل حالٍ وفق إعلان أصحاب المشروع، وتوسع مستقبلي مخطط، وهدف بعيد يعتمد على التنفيذ والطلب والتمويل S09.

يزيد هذا الخبر السؤال السعودي عمقاً. فوجود معالجات من «إيه إم دي» (AMD) في خدمة مدفوعة محتملة، بدل اقتصار الحديث على الشراء من «إنفيديا» (NVIDIA)، يفتح باباً لمقارنة مجموع التكلفة (Total Cost of Ownership; TCO)، والتوافق البرمجي (Software Compatibility)، وكلفة نقل النموذج (Model Portability)، وأداء التطبيقات الفعلي عبر عائلات عتاد مختلفة. لكنه لا يبرهن أن منصة واحدة أرخص أو أكثر ربحية من الأخرى، ولا يكشف عقود العملاء أو مستويات الاستخدام أو الهامش الصافي. هذه تحتاج بيانات غير واردة في الإعلان.

وقبل ذلك بعام، أعلنت «هيومين» (HUMAIN) إطلاق تطبيق محادثة عربي يعتمد نموذج «علام 34 مليار» (ALLaM 34B)، بخصائص تشمل إدخال الصوت العربي والبحث واستعمال العربية والإنجليزية. الخبر لا يخبرنا كم رمزًا خُدم، ولا إيرادًا تحقق من كل جلسة؛ لكنه ينقل صورة المركز الحاسوبي من مجسم طاقة إلى استعمال ممكن على هاتف مواطن أو داخل عمل مؤسسة. وهذا هو الانتقال الاقتصادي الحاسم: لا توجد سوق للحواسيب لمجرد أنها تفكر؛ توجد سوق لأن إنسانًا أو منشأةً وجدا عملاً يستحق أن يوكلا إليها وينفقا عليه S10.

فإذا أرادت السعودية أن تتقدم من شراء البنية إلى امتلاك موقع تنافسي في صناعة الاستدلال (Inference Industry)، فليس عليها أن تختار بين «أكبر مركز» و«أرخص رمز» كأنهما نهاية التفكير. عليها اكتشاف أنواع العمل ذات القيمة العالية التي تستطيع تسليمها بجودة متوقعة وبكلفة قابلة للانخفاض مع الزمن. قد تكون سلاسل مطورة للغة العربية، أو نماذج متخصصة مع اختبارات دقة، أو استضافة معالجة عالية الخصوصية، أو خدمات تطوير نماذج ذات زمن استجابة مضمون. هذا أمثلة على فئات اقتصادية ضمن محور الحوسبة، لا إعلانًا بأنها قائمة الآن أو أنها كلها رابحة.

7. اختبار واحد قد يجعل خمسة معالجات متشابهة مختلفة تمامًا

للإعلانات التقنية قدرة لطيفة على تحويل الاختلاف الكبير إلى كلمة واحدة: «الأداء». لكن محرك استدلال يخدم خمسة أسئلة قصيرة في وقت واحد ليس المحرك نفسه عندما يحمل ألف محادثة طويلة تتبادل استخدام السياق. سرعة الرمز الأول (Time to First Token; TTFT) غير المسافة الزمنية بين الرموز (Inter-Token Latency; ITL)، وكلاهما غير عدد الرموز المعالجة في الثانية (Throughput). وقد يربح جهاز مسابقة الرموز لكنه يتأخر في أول جواب، فيخسر تطبيقًا تهمة الاستجابة الفورية؛ ويخسر جهاز آخر ترتيب السرعة الفردية لكنه يربح في الموازنة العالية.

صدر في 16 سبتمبر 2026 تحديث اختبار «إم إل بيرف للاستدلال» (MLPerf Inference v6.1) من مؤسسة «إم إل كومونز» (MLCommons). تضمّن اختبارات جديدة أقرب إلى المهام المتعددة الخطوات، ونتائج أنظمة جديدة تُراجع وفق قواعد الاختبار. وهذه إضافة معرفية مهمة، لكنها لا تمنح شهادة ربح مالي لأي مصنع. فالاختبار يقيس أداء تشغيل وفق قواعد محددة، وليس الثمن الذي يحققه مشغل الحوسبة من زبائنه، ولا تمويله، ولا الضريبة، ولا تكلفة بقاء الأجهزة دون استخدام S11.

ليتخيل القارئ منافسة حقيقية بين جهازين من الجيل نفسه، لا يعرف أيهما أفضل بعد:

- الجهاز الأول يخرج أربعة آلاف رمز في الثانية في حمل معياري، لكن تكلفة التشغيل الشهرية لكل عميل عالية لأنه يحتاج إلى حجز سعة كبيرة لضمان سرعة الرمز الأول.
- الثاني يخرج ثلاثة آلاف وخمسمئة فقط في ذلك الحمل، لكنه يشارك ذاكرة السياقات، ويشغل دفعات متغيرة بكفاءة، ويستفيد من طلبات المستخدمين المتكررة؛ فيتيح لنفسه خدمة عدد أكبر من العملاء ضمن الاتفاق الزمني الذي يدفعون مقابلته.

هذه أرقام توضيحية مختلفة لتبيان طريقة المقارنة، وليست قياساً لجهازين منشورين. لا نستطيع إعلان الفائز إلا إذا ثبتنا نموذج الذكاء، وطول المدخلات والمخرجات، وجودة الإجابة، والنسب الفعلية للطلبات، وحدود التأخير، وكلفة الكهرباء والتبريد والتشغيل، ونسبة العمل المفوتر. عندها تصبح المقارنة «تكلفة الطلب الصحيح والمقبول» (Cost per Accepted Task) أو «هامش الخدمة المؤهلة» (Margin per SLA-Compliant Task) أكثر اتصالاً بالحياة الاقتصادية من مقارنة الأجهزة بحسابات معزولة.

هنا يخرج السؤال من صناديق المعالجات إلى مكتب مدير استثمار: أي مجموعة من أحمال العمل تصلح أصلاً لهذا الجهاز؟ إذا كان العملاء يريدون توليداً متتابعاً طويلاً، أصبحت الذاكرة ونطاق نقلها قيداً. وإذا كان معظم الاستخدام بحثاً في وثائق مؤسسية مشتركة، فقد يرتفع أثر إعادة استخدام السياق. وإذا كان العملاء يحتاجون تدريب نماذج جديدة، اختلف توزيع القيمة والحساسية لتبديل الأجيال. ليس من الذكاء أن نضع كل هذه الفواتير في آلة حساب اسمها «سعر الساعة» ثم نندهش من اختلاف النتائج.

8. الحيلة التي لا تصلح بها شركة الكهرباء عقداً للبيع

في عالم الحوسبة ثلاث سرعات يصعب جمعها في لوحة واحدة: سرعة تصنيع الجيل، وسرعة تركيب القدرة الكهربائية، وسرعة اكتشاف طلب يدفع مقابل النتيجة. الأولى تتحرك في دورات المنتجات والأبحاث؛ والثانية في إنشاءات وشبكات وتوريد؛ والثالثة في تجارب مؤسسات قد تبدأ متحمسة ثم تجد أنها لا تحتاج الاستدلال طوال الوقت. ومن هذه الفجوة تنشأ قيمة «القدرة الاختيارية» (Real Option): أن تستطيع تأجيل شحنة أو تعديل نوع الأحمال أو استئجار السعة مرحلياً بدل تحويل كل طموح مستقبلي إلى فاتورة آنية.

لننظر إلى العقد الذي لا تظهر صورته في إعلانات مراكز البيانات. تقول «كور ويف» (CoreWeave) إن كثيراً من عقودها متعددة السنوات من نوع «خذ أو ادفع» (Take-or-Pay)، وبلغ متوسط المدة المرجح لعقودها الملتزم بها نحو خمس سنوات في نهاية 2025. ومعناه في عقودها الموصوفة أن العميل يدفع مقابل قدرة ملتزم بها وفق الشروط حتى إن لم يستخدمها بالكامل. هذه آلية اقتصادية قادرة على نقل جزء من خطر انخفاض الاستخدام بعيداً عن المشغل؛ ولذلك ليس من العلمي توقع أثر تخفيض أسعار واجهات الذكاء على إيراده كما لو أن كل خدماته تُباع فوراً وفق سعر لحظي مفتوح S01.

العقد، مع ذلك، ليس درعاً سحرياً ولا لعبة تمحو الزمن. يختلف أثره بحسب قدرة الشركة على التسليم، وحقوق العميل، وشروط الأسعار والتجديد، وملاءة الأطراف، ومواعيد دفع الديون. إذا فقد عميل كبير حاجته الاقتصادية إلى القدرة المحجوزة، فقد يستمر بالدفع حتى انتهاء العقد ثم يفاوض بقسوة عند التجديد. وإذا وجد مركز حوسبة فرصة مربحة بسعر أعلى، قد لا تكون لديه حرية نقل قدرة سبق أن تعهد بها. الهامش الذي تمنحه الاستمرارية يمكن أن يصبح تكلفة فرصة ضائعة عند انقلاب الظروف. لكننا لا نملك تفاصيل العقود السعودية، فلا يصح إسقاط معادلة شركة أمريكية عليها تلقائياً.

إن سر الحكم على عائد المعالج في توقيت الفوائض والعجوزات، وليس فقط مجموع التدفقات. عند قرض يستحق قبل نضج الطلب يمكن أن يواجه المشروع ضغطاً نقدياً رغم قيمة حالية صافية موجبة لاحقاً. وعند عقد طويل بسعر ثابت قد يحمي التدفقات من هبوط السعر لكنه يمنع المشاركة الكاملة في صعوده. وعند تحسن برمجي قوي قد يرتفع هامش الوحدة ويهبط إيراد الوحدة في آن. لا يجوز تحويل هذه التبادلات إلى قائمة «مخاطر حوكمة»، فكل واحدة منها علاقة مالية قابلة للحساب: جدول سداد، منحنى طلب، حد استهلاك، موعد جيل تقني، ونافذة قدرة كهربائية.

9. فكرة استراتيجية: لا تملك المعالج بوصفه قطعة؛ امتلك منحنى تعلمه الاقتصادي

قد يكون أعظم أصل لدى مشغل حوسبة في 2030 شيئاً لا تصل إليه الكاميرا في جولة مركز البيانات: سجل تاريخي دقيق يربط نوع الحمل بما استهلكه فعلاً وبما دفع العميل من أجله وبما قبله من جودة. ليس سجلاً تجميليًا لعدد الاستدعاءات، بل معرفة تعلمت، شهرًا بعد شهر، أين تبدد الزمن وأين تكوّن الهامش.

أسمي هذا مقترحًا بحثيًا «خريطة العمر التجاري للحوسبة» (Compute Commercial Life Map). لكل فئة حمل عمل ندون خمس ساعات مختلفة: زمن إعداد النموذج للعتاد، زمن الانتظار قبل أول مخرج، زمن إنتاج مخرج مقبول، زمن بقاء السعر قابلاً للدفاع تنافسيًا، وزمن الاسترداد النقدي لنصيبها من الاستثمار. ليست الساعات متساوية، ولا تصلح كل واحدة أن تختزل في متوسط وطني واحد. قد يكون عمر الشريحة الفيزيائي ست سنوات، وعمر مزيتها السعرية ثمانية أشهر، وعمر فائدتها في نموذج أصغر أطول بكثير، وعمر عقد يشغلها مدة محددة لا علاقة له بهذه الأعمار الثلاثة.

إن بائع آلة يُقنعك عادةً بأفضل اختبار. أما مدير الاستثمار المخضرم فيريد منحنى انطفاء الأفضلية الاقتصادية (Economic Advantage Decay Curve): كم شهرًا تبقى الخدمة أعلى قيمة من بديلها قبل أن يحتاج إلى خفض السعر أو تبديل نوع العمل؟ نحتاج إلى قياس ذلك بحسب دفعات المعدات، لا حسب شعار جيل المعالج. عندئذ يمكن مقارنة ثلاثة أفعال لا تظهر في عنوان السباق: تحسين البرنامج، إعادة تسعير الخدمة، أو تغيير العملاء الذين تناسبهم القدرة. وربما يكون قرار عدم شراء دفعة إضافية أذكى قرار حوسبي في العام، مع أن الإعلان عنه لا يصلح صورةً افتتاحية براقعة.

يوجد هنا رهان مضاد للتشاؤم يستحق الاحترام: انخفاض تكلفة الاستدلال قد يخلق تطبيقات لم تكن قابلة للبيع أصلاً، فيرفع إجمالي السوق ويعزز الربحية على الرغم من تراجع السعر. وقد تمتص الأعمال المعقدة الجديدة كل التحسن، بدل إطلاق حرب أسعار. كذلك قد تستفيد السعودية من توفر طاقة وقدرة استثمارية وأسواق إقليمية وتطبيقات عربية. لكنها شروط وطرق اختبار، وليست نتائج تُستخلص من حجم مشروع معلن. ومن دون مشاهدة الطلب المدفوع وهامش الوحدة والعقود المنفذة، تبقى الحكايتان ممكنتين.

10. خمسة أشياء تستحق أن تقرأها قبل الإعلان القادم

بوسع الباحث أو المسؤول الاستثماري أن يبدأ غداً، إن امتلك بيانات تشغيلية وتجارية مخولة ومجمعة، من غير انتظار مؤشر عبقرى غير موجود:

ما نقيسه	السؤال الذي يكشفه	الخطأ إذا اكتفينا بالرقم الأسهل
إيراد العمل المقبول (Revenue per Accepted Task)	هل يدفع العميل أكثر من التكلفة الكاملة لتنفيذ مهمة ناجحة؟	الاكتفاء بثمن مليون رمز مهما اختلفت المهمة.
السعة الفعالة ضمن مستوى خدمة محدد (SLA- Constrained Throughput)	كم طلباً يمكن إنجازه بجودة وتأخير متفق عليهما؟	استعمال الحد الأقصى المعمل للمعالج.
نسبة الإيراد المحجوز تعاقدياً (Contracted Revenue Share)	ما جزء التدفق الذي لا يعتمد على السعر الفوري، وتحت أي شروط؟	مساواة إعلان عقد بتحصيل نقدي أو هامش مضمون.
مدة الاسترداد تحت هبوط السعر (Price-Adjusted Payback)	هل يستعيد المشروع رأس ماله في الزمن الذي تصمد فيه ميزته التجارية؟	الاكتفاء بعمر الإهلاك المحاسبي.
انتشار عوائد الأحمال (Workload Margin Distribution)	هل يخفي المتوسط أحمالاً تخسر وأخرى تمولها؟	عرض إشغال شامل جميل يطمس اختلاف التطبيقات.

لنفترض اختباراً معلناً في السوق السعودية: مزودان يعملان على أعباء متقاربة مجهولة الهوية، ويتنافسان على تقديم خدمة عربية معروفة الجودة خلال حد زمني واحد. تُقاس تكلفتها الكاملة لكل مهمة مقبولة، ثم تتكرر التجربة عند تغيير حجم الطلب وتوزع أطوال المحادثات ونسبة إعادة استخدام السياق. يضاف اختبار اقتصادي: ما أثر خفض سعر الخدمة عشرين بالمئة في العملاء المكتسبين، وفي الطلبات المكتملة، وفي الربح؟ لا نحتاج كشف تفاصيل منشآت حيوية أو أسرار عملاء؛ يكفي نشر المقارنات المجمعة وتعريفات القياس. والنتيجة غير مضمونة سلفاً: قد ينتصر الجهاز الأحدث، أو البرنامج الأفضل، أو العمل الأنسب، أو العقد المتزن.

هذه هي الفرضية التي أقترحها في هذه المقالة: «التفوق الوطني في حوسبة الذكاء الاصطناعي لا ينشأ من امتلاك أكبر قدرة مجردة، بل من سرعة تحويل التحسينات التقنية إلى أعمال نافعة مدفوعة قبل أن تتآكل أسعارها». جملة تصلح للاختبار؛ فإذا لم نجد علاقة بعد ضبط أحمال العمل والتكلفة وشروط العقود، ينبغي لنا إسقاطها أو تضيقها. لا أريد أن أؤسس سوقاً من استعارة جميلة.

11. ماذا سيقول لنا عام 2030، لو أحسنّا السؤال اليوم؟

في تصور أول، يتسع الطلب على الاستدلال المعقد أكثر من انخفاض سعره؛ يصبح إطلاق جيل عتاد جديد مناسبة لفتح تطبيقات لا تستقيم على الجيل السابق، وتستفيد القدرات السعودية من سوق إقليمية أكبر. في تصور ثانٍ، تُنتج البرمجيات والحواسيب الجديدة وفرة هائلة، لكن معظم القيمة ينتقل إلى مطوري التطبيقات والمستخدمين؛ يتحول

مشغلو مراكز الحوسبة إلى منافسة محتدمة على السعر والتكاليف. وفي تصور ثالث، يتعايش القطاعان: قدرة فائقة لأعمال محدودة ذات أثمان مرتفعة، وحوسبة عامة أصبحت سلعة منخفضة الهامش. هذه مسارات مشروطة غير مرجحة احتماليًا؛ ليست ثلاثة تنبؤات محسوبة.

ماذا نراقب بين 2026 و2030؟ تغير الأسعار المحققة لا المنشورة وحدها، النسبة الفعلية للعقود الملتزمة، تركيب المدخلات والمخرجات والمخبا، أداء الخدمة ضمن حدود زمنية موحدة، فارق تكلفة المهام بين أجيال العتاد على أحمال واحدة، وحصة الإيراد الآتي من عملاء خارج السوق المحلية. وإذا ارتفع حجم الرموز أسرع من إيرادها فلا نعلن كارثة؛ نفحص الهامش، ثم استرداد الأصل، ثم تكلفة اكتساب العملاء. وإذا ارتفع الإيراد فلا نعلن نصرًا؛ نفحص كم ابتلع رأس المال والتبريد والشبكات منه.

مدير المصنع القديم كان يستمع إلى إيقاع آلاته ويعرف أن السكون خسارة محتملة. مدير مصنع الذكاء الحديث يحتاج إلى شجاعة مغايرة: أن يرى المعالج ينتج أكثر، ثم يسأل بأدب شديد وصرامة أشد عمّن دفع مقابل هذا الإنتاج، وعن العمل الذي بقي غير مفوتر، وعن عميل كان يستطيع استخدام جهاز أصغر بالحساب نفسه.

فالآلة الحديثة قد تفوز في كل اختبار سرعة وتخسر اجتماع التسعير التالي. وقد يخسر جيل قديم ترتيب الأداء ويظل يربح لأنه ينفذ عملاً يعرف زبائنه جيدًا. لا تناقض بينهما. تناقضنا نحن حين نخلط تقدّم العلم بتقدّم الفاتورة.

والسؤال الذي ينبغي أن يبقى في ذهن القارئ ليس: كم سيبلغ عدد المعالجات في السعودية؟ بل: ما الذي ستصبح السعودية أبرع من غيرها في حسابه، ومن يحتاج ذلك الحساب بما يكفي ليدفع ثمنه حتى بعد وصول المعالج التالي؟

المصادر والأعمال المحال إليها

1. [CoreWeave, 2025 Form 10-K, SEC](#), العمر الإنتاجي والمعلومات التعاقدية، نُشر 2026.
2. [CoreWeave, Form 10-Q, period ending 30 June 2026, SEC, Note 2](#) و RPO.
3. [NVIDIA, Rubin platform announcement, 5 January 2026](#).
4. [OpenAI, Introducing GPT-6 Sol and Luna, September 2026, pricing and caching sections](#).
5. [Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention, SOSP 2023, DOI 10.1145/3600006.3613165](#).
6. [Yu et al., Orca: A Distributed Serving System for Transformer-Based Generative Models, OSDI 2022](#).
7. [NVIDIA, TensorRT-LLM, Disaggregated Serving technical documentation](#).
8. [William Stanley Jevons, The Coal Question, 1865, digitized historical edition](#).

9. AMD, Cisco and HUMAIN, joint production and expansion announcement, 31

August 2026

10. Saudi Press Agency, HUMAIN Chat powered by ALLaM 34B, 25 August 2025

11. MLCommons, MLPerf Inference v6.1 Results, 16 September 2026

12. Reproducible original **synthetic** model, assumptions and 27 scenario outputs؛ انظر

جدول النتائج والافتراضات.

ملاحظة تحريرية: الأرقام المالية المنشورة حقائق من مصادرها مع نسبتها؛ نتائج المحاكاة ليست بيانات عن شركات أو خدمات سعودية. المؤسسات المذكورة لم تشارك في إعداد هذا العمل أو مراجعته.